Statistica Sinica **20** (2010), 405-421

RELEASING MULTIPLY-IMPUTED SYNTHETIC DATA GENERATED IN TWO STAGES TO PROTECT CONFIDENTIALITY

Jerome P. Reiter and Jörg Drechsler

Duke University and Institute for Employment Research

Abstract: To protect the confidentiality of survey respondents' identities and sensitive attributes, statistical agencies can release data in which confidential values are replaced with multiple imputations. These are called synthetic data. We propose a two-stage approach to generating synthetic data that enables agencies to release different numbers of imputations for different variables. Generation in two stages can reduce computational burdens, decrease disclosure risk, and increase inferential accuracy relative to generation in one stage. We present methods for obtaining inferences from such data. We describe the application of two stage synthesis to creating a public use file for a German business database.

Key words and phrases: Confidentiality, disclosure, multiple imputation, synthetic data.

1. Introduction

Many national statistical agencies, survey organizations, and researchers–henceforth called agencies–disseminate microdata, i.e., data on individual units in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases and improvements in record linkage technologies have increased the risk of disclosure to the point where most agencies alter microdata before release (Reiter (2004a)). For example, agencies globally recode variables, such as releasing ages in five year intervals or top-coding incomes above 100,000 as "100,000 or more"; they swap data values for randomly selected units; or, they add random noise to data values. When applied with high intensity, these strategies reduce the utility of the released data, making some analyses impossible and severely distorting the results of others. They also complicate secondary analyses: adjusting inferences for data alterations may be beyond some public data users' statistical capabilities.

An alternative approach to disseminating public use data was suggested by Rubin (1993): release multiply-imputed, synthetic data sets. Specifically, he

proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these data sets to the public. A related approach was suggested by Fienberg (1994). These are called *fully synthetic* data sets. Releasing fully synthetic data can protect confidentiality, since identification of the sampled units and their sensitive data is very difficult when the released data are not the original records and do not contain collected values. Furthermore, with appropriate synthetic data generation and the inferential methods developed by Raghunathan, Reiter and Rubin (2003) and Reiter (2005c), users can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of fully synthetic data are described by Rubin (1993), Little (1993), Fienberg, Makov and Steele (1998), Raghunathan et al. (2003), and Reiter (2002, 2005b).

Some agencies have adopted a variant of Rubin's original approach, suggested by Little (1993): release data sets comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic* data sets. For example, the U.S. Federal Reserve Board protects data in the Survey of Consumer Finances by replacing large monetary values with multiple imputations (Kennickell (1997)). The U.S. Bureau of the Census (Abowd and Woodcock (2001, 2004)) protects data in longitudinal data sets by replacing all values of sensitive variables with multiple imputations and leaving other variables at their actual values. Little, Liu and Raghunathan (2004) present an algorithm, named SMIKe, for simulating multiple values of key identifiers for selected units. Partially synthetic, public use data are being developed for the U.S. for the Survey of Income and Program Participation, the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Community Survey group quarters data.

Partial synthesis is appealing because it promises to maintain the primary benefits of full synthesis–protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software–with decreased sensitivity to the specification of imputation models. Valid inferences from partially synthetic data sets can be obtained using the methods developed by Reiter (2003, 2005c), whose rules for combining point and variance estimates differ from those of Rubin (1987) and also from those of Raghunathan et al. (2003). Methods for handling missing data simultaneously with partially synthetic data are developed in Reiter (2004b). Other illustrations of partially synthetic data include Reiter (2005d) and Mitra and Reiter (2006).

Fully and partially synthetic data have a key difference. Each fully synthetic data set comprises independent samples of records off the frame, whereas

each partially synthetic data set comprises the original records. Essentially, full synthesis simulates repeated sampling of the population, and partial synthesis modifies collected values for the original records.

In this article, we present a two-stage approach to generating fully and partially synthetic data in which agencies impute some variables only a few times and other variables many times. Two stage synthesis can have advantages over one-stage synthesis. In some settings, it reduces disclosure risks while increasing data usefulness. For example, agencies may want to release only a few imputed values of quasi-identifiers or sensitive variables, since intruders can use information from multiple data sets to refine guesses of the true values (Little et al. (2004), Reiter (2005d), and Mitra and Reiter (2006)), but they may want to release large numbers of imputations for other variables to drive down the variance introduced by imputation. In other settings, it reduces the labor needed to generate synthetic data. This is the case for the two-stage synthesis of the public release data for the German Institute for Employment Research (IAB) Establishment Panel, which is described in Section 2. A related approach, called nested multiple imputation (Shen (2000), Harel and Schafer (2003), and Rubin (2003)), has been used to reduce labor in the context of imputation for missing data.

The remainder of this article is organized as follows. Section 2 motivates the usefulness of two-stage synthetic data for reducing disclosure risks or decreasing agencies' labor. Section 3 presents methods for obtaining inferences from two-stage synthetic data. Section 4 illustrates the performance of these methods via simulation studies. Section 5 concludes with general remarks about synthetic data.

2. Motivation for Two-Stage Synthesis

We first review evidence from the literature on the implications for disclosure risk and inferential accuracy of releasing many synthetic data sets. Two-stage synthesis allows agencies to compromise on the risk-accuracy trade-off. We then describe the synthesis of data from the IAB Establishment Panel, for which one-stage synthesis demands too high labor cost.

2.1. Implications of releasing many synthetic data sets

From the perspective of the data analyst, there are benefits when agencies release a large number of synthetic data sets. The variability in point estimates decreases with the number of replicates. The reduction can be substantial when many values are synthesized. For example, Reiter (2002) finds a 30% increase in the variance of survey-weighted estimates of population means when dropping from one hundred to five fully synthetic data sets. Reiter (2003) finds nearly a 100% increase in the variance of regression coefficients when going from fifty to

two partially synthetic data sets in which all values of a dependent variable are replaced with imputations. Increasing the number of replicates also reduces the variability in estimators of variance. This variability can be large when many values are synthesized; in fact, for fully synthetic data, Reiter (2005b) finds that some variance estimators computed with ten fully synthetic data sets are so poor as to be essentially worthless. Those variance estimators have acceptable properties with one hundred replicates. The incremental benefits become minimal as the number of replicates gets large.

From the perspective of the agency, there are risks to releasing a large number of synthetic data sets. Increasing the number of replicates provides more information for intruders to estimate the original data values. To illustrate this, we extend the partial synthesis done by Mitra and Reiter (2006), which used the 1987 U.S. Survey of Youth in Custody. The survey interviewed youths in juvenile facilities about their family background, previous criminal history, and drug and alcohol use. The sample contains 2,621 youths in 50 facilities. Mitra and Reiter (2006) consider facility membership to be potentially identifying information. Therefore, they generated new facility identifiers for all youths. This was done by (i) fitting multinomial regressions of facility identifiers on the survey variables, (ii) drawing new values of parameters for the regressions and computing the resulting predicted probabilities, and (iii) simulating new identifiers from the multinomial distributions based on the predicted probabilities. To assess disclosure risk, they assumed that the intruder uses the mode of each youth's multiply-imputed facility as the best guess of the youth's actual facility. When no unique mode exists, they randomly select one value. We followed the same procedures for different numbers of synthetic data sets. With three replicates, approximately 17% of intruders' guesses were correct. With ten replicates, this increases to 20%. With fifty replicates, this increases to 24%. While perhaps not alarming, the increasing identification rates certainly would push agencies to minimize the number of imputations of facilities.

For fully synthetic data, there has been little work on the impacts on disclosure risk of releasing many replicates. In part, this is because identification disclosure risks are low for fully synthetic data. Each data set contains different samples of records, and all survey variables are synthesized. However, the risks are not zero. When the imputation models are highly detailed, the imputations could reproduce combinations of quasi-identifiers for real records. Intruders might interpret this to mean that real-data records with those characteristics were in the original sample, which could result in identification disclosures if some of those records are unique in the population. This risk could be magnified when releasing multiple synthetic data sets, because (i) there are several opportunities to impute such records, and (ii) there could be repetitions of realistic synthetic records that might strengthen the intruder's confidence that a similar real record was in the original data.

Ideally, when considering the release of public use data, the agency balances confidentiality protection and inferential accuracy; see, for example, Duncan, Keller-McNulty and Stokes (2001), Reiter (2005a), Gomatam, Karr, Reiter and Sanil (2005), and Karr, Kohnen, Oganian, Reiter and Sanil (2006). Confidentiality concerns often trump accuracy concerns. With one-stage synthesis, favoring confidentiality over accuracy could lead agencies to release few replicates. With two-stage synthesis, agencies can compromise on the risk-accuracy trade-off. Agencies can release few imputations of quasi-identifiers or other confidential variables to reduce disclosure risks, and release many imputations of other variables to enable analysts to improve precision for analyses involving those variables.

2.2. Synthesis of the IAB establishment panel

The IAB Establishment Panel, conducted since 1993, contains detailed information about German firms' personnel structure, development, and policy. Considered one of most important business panels in Germany, there is high demand for access to these data from external researchers. Because of the sensitive nature of the data, researchers desiring direct access to the data have to work on site at the IAB. Alternatively, researchers can submit code for statistical analyses to the IAB research data center, whose staff run the code on the data, and send the results to the researchers. To help researchers develop code, the IAB provides remote access to a publicly available "dummy data set" with the same structure as the Establishment Panel. The dummy data set comprises random numbers generated without attempts to preserve the distributional properties of the variables in the Establishment Panel data. For all analyses done with the genuine data, researchers can publicize their analyses only after IAB staff check for potential violations of confidentiality.

Releasing public use files of the Establishment Panel would allow more researchers to access the data with fewer burdens, stimulating research on German business data. It also would free up staff time from running code and conducting confidentiality checks. Because there are so many sensitive variables in the data set, standard disclosure limitation methods like swapping or microaggregation would have to be applied with high intensity, which would severely compromise the utility of the released data. Therefore, the IAB decided to develop synthetic data, specifically (at this stage) fully synthetic data.

Each synthetic data set comprises establishments sampled from the sampling frame for the Establishment Panel. Records are sampled according to the design of the Establishment Panel–stratifying by region, establishment size, and industry–to take advantage of the efficiency gained by the original stratification. Let X be the variables corresponding to the stratum indicators.

Values of the Establishment Panel survey variables, Y_b, are imputed for all establishments in the synthetic data samples. These models are developed as follows. First, for all records in the original panel, establishment-level data, Y_a, are obtained from the German Social Security Data (GSSD). The GSSD contains information on individuals covered by social security, including data on their employer such as demographic characteristics and average wages of its employees. The employers are identified by the establishment identification numbers used in the Establishment Panel, which enables direct matching between the two data sources. Second, a statistical model relating Y_b to (X, Y_a) is estimated using the data from the original panel. Third, for each synthetic sample, the newly drawn establishments are matched to the GSSD and values of Y_a are appended to the synthetic data. Fourth, values of Y_b are simulated from $f(Y_b|X, Y_a)$, using the X and the appended values of Y_a for the new establishments. After the imputation, all variables in Y_a are deleted for confidentiality reasons. The result is a synthetic data set that mimics the structure of the Establishment Panel, comprising the stratification indicators X and the imputed survey variables Y_b.

Previous research has shown that releasing large numbers of fully synthetic data sets improves synthetic data inferences (Reiter (2005b)). The usual advice from multiple imputation for missing data–release five multiply-imputed data sets–tends not to work well for fully synthetic data because the fractions of "missing" information are large. Following Reiter (2005b), the IAB desired to generate and release one hundred fully synthetic data sets. However, doing so requires matching to the GSSD one hundred times and imputing Y_b for each matched sample. These are very labor intensive tasks; the matching has to be checked and corrected if necessary each time, and the matched data need to be transferred to different software platforms to impute Y_b. Furthermore, each matched data file is re-configured manually to implement the imputation routines.

This led the IAB synthesis team to adopt a two-stage approach to synthesis. Only ten synthetic samples are drawn, thus requiring only ten iterations of matching and data processing to obtain Y_a. For each sample, Y_b is imputed another ten times, resulting in one hundred data sets. This two-stage method reduces the labor by a factor of ten while allowing the release of one hundred data sets containing information about Y_b, as opposed to only ten. For more details about the imputation models in the synthesis, based on the sequential multivariate regression imputation strategy of Raghunathan, Lepkowski van Hoewyk and Solenberger (2001), see Drechsler, Dundler, Bender, Rässler and Zwick (2008).

The ten sets of Y_b for each sample are correlated. Existing inferential methods for synthetic data do not account for this correlation. We present inferential methods that do so for both full and partial two-stage synthesis in Section 3. The methods are presented assuming all variables are released, but they apply when

some variables are suppressed as in the synthesis of the Establishment Panel. The methods also assume for generality that (Y_a, Y_b) is known only for the sampled records.

3. Inferences with Two-Stage Synthetic Data

For a finite population of size N, let $I_l = 1$ if unit l is included in the survey, and $I_l = 0$ otherwise, where $l = 1, \ldots, N$. Let $I = (I_1, \ldots, I_N)$, and let the sample size $s = \sum I_l$. Let X be the $N \times d$ matrix of sampling design variables, e.g. stratum or cluster indicators or size measures. We assume that X is known approximately for the entire population, for example from census records or the sampling frame(s). Let Y be the $N \times p$ matrix of survey data for the population. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $s \times p$ sub-matrix of Y for all units with $I_l = 1$, where Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let R be an $N \times p$ matrix of indicators such that $R_{lk} = 1$ if the response for unit l to item k is recorded, and $R_{lk} = 0$ otherwise. The observed data is thus $D_{obs} = (X, Y_{obs}, I, R)$.

3.1. Fully synthetic data

Let Y_a be the values simulated in stage 1, and let Y_b be the values simulated in stage 2. The agency seeks to release fewer replications of Y_a than of Y_b, yet do so in a way that enables the analyst of the data to obtain valid inferences with standard complete data methods. To do so, the agency generates synthetic data sets in a three-step process. First, the agency fills in the unobserved values of Y_a by drawing values from $f(Y_a \mid D_{obs})$, creating a partially completed population. This is repeated independently m times to obtain $Y_a^{(i)}$, for $i = 1, \ldots, m$. Second, in each partially completed population defined by nest i, the agency generates the unobserved values of Y_b by drawing from $f(Y_b \mid D_{obs}, Y_a^{(i)})$, thus completing the rest of the population values. This is repeated independently r times for each nest to obtain $Y_b^{(i,j)}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, r$. The result is $M = mr$ completed populations, $P^{(i,j)} = (D_{obs}, Y_a^{(i)}, Y_b^{(i,j)})$, where $i = 1, \ldots, m$ and $j = 1, \ldots, r$. Third, the agency takes a simple random sample of size n_{syn} from each completed population $P^{(i,j)}$ to obtain $D^{(i,j)}$. These M samples, $D_{syn} = \{D^{(i,j)} : i = 1, \ldots, m; j = 1, \ldots, r\}$, are released to the public. Each released $D^{(i,j)}$ includes a label indicating its value of i, i.e. an indicator for its nest.

The agency can sample from each $P^{(i,j)}$ using designs other than simple random samples, such as the stratified sampling in the IAB Establishment Panel synthesis. A complex design can improve efficiency and ensure adequate representation of important sub-populations for analyses. When synthetic data are generated using complex samples, analysts should account for the design in inferences, for example with survey-weighted estimates. One advantage of simple

random samples is that analysts can make inferences with techniques appropriate for simple random samples.

The agency could simulate Y for all N units, thereby avoiding the release of actual values of Y. In practice, it is not necessary to generate completed-data populations for constructing the $D^{(i,j)}$; the agency need only generate values of Y for units in the synthetic samples. The formulation of completing the population, then sampling from it, aids in deriving inferential methods.

Let Q be the estimand of interest, such as a population mean or a regression coefficient. The analyst of synthetic data seeks $f(Q|D_{syn})$. The three-step process for creating D_{syn} suggests that

$$f(Q|D_{syn}) = \int f(Q|D_{obs}, P_{syn}, D_{syn}) f(D_{obs}|P_{syn}, D_{syn}) \times f(P_{syn}|D_{syn}) dD_{obs} dP_{syn}, \tag{3.1}$$

where $P_{syn} = \{P^{(i,j)} : i = 1, \ldots, m; j = 1, \ldots, r\}$. As in other applications of multiple imputation approaches, we find each component of this integral by assuming that the analyst's distributions are identical to those used by the agency for creating D_{syn}. We also assume that the sample sizes are large enough to permit normal approximations for these distributions. Thus, we require only the first two moments for each distribution, which can be derived using standard large sample Bayesian arguments. Diffuse priors are assumed for all parameters.

Integration can be carried out numerically, as we describe in the Supplement to the on-line version of the article (`http://www.stat.sinica.edu.tw/statistica`). Here, we present an approximation that can be easily computed by analysts using D_{syn}. Its derivation also is in the Supplement. For all (i, j), let $q^{(i,j)}$ be the estimate of Q, and let $u^{(i,j)}$ be the estimate of the variance associated with $q^{(i,j)}$. The $q^{(i,j)}$ and $u^{(i,j)}$ are computed based on the design used to sample from $P^{(i,j)}$. Note that when $n_{syn} = N$, the $u^{(i,j)} = 0$. Let $\bar{q}_r^{(i)} = \sum_j q^{(i,j)}/r$, and $\bar{q}_M = \sum_i \bar{q}_r^{(i)}/m$. Let $b_M = \sum_i (\bar{q}_r^{(i)} - \bar{q}_M)^2/(m-1)$, and $w_r^{(i)} = \sum_j (q^{(i,j)} - \bar{q}_r^{(i)})^2/(r-1)$. Finally, let $\bar{u}_M = \sum_{i,j} u^{(i,j)}/(mr)$.

For large m and r, we approximate $f(Q|D_{syn})$ by a normal distribution with $E(Q|D_{syn}) = \bar{q}_M$ and $Var(Q|D_{syn}) = (1 + m^{-1})b_M + (1 - 1/r)\bar{w}_M - \bar{u}_M = T_f$. For modest m and r, we obtain inferences by using a t-distribution, $(\bar{q}_M - Q) \sim t_{\nu_f}(0, T_f)$. The degrees of freedom, ν_f, are

$$\nu_f = \left(\frac{((1 + 1/m)b_M)^2}{(m-1)T_f^2} + \frac{((1 - 1/r)\bar{w}_M)^2}{(m(r-1))T_f^2} \right)^{-1}.$$

The degrees of freedom are derived by matching the first two moments of T_f to a chi-squared distribution with ν_f degrees of freedom, as shown in the Supplement.

It is possible that $T_f < 0$, particularly for small values of m and r. Generally, negative values of T_f can be avoided by making n_{syn} or m and r large. To adjust for negative variances, one approach is to use the always positive variance estimator, $T_f^* = T_f + \lambda \bar{u}_M$, where $\lambda = 1$ when $T_f \leq 0$ and $\lambda = 0$ when $T_f > 0$. When $T_f < 0$, using ν_f is overly conservative, since T_f^* tends to be conservative when $\lambda = 1$. To avoid excessively wide intervals, analysts can base inferences on t-distributions with degrees of freedom $\nu_f^* = \nu_f + \lambda\infty$.

Rather than t-approximations, analysts willing to use Monte Carlo methods can simulate $f(Q|D_{syn})$ directly; see the Supplement. This can be done for any (m, r) and avoids adjustments for negative variance estimates. Agencies could disseminate software routines to facilitate such simulations.

3.2. Partially synthetic data

We assume that $Y_{inc} = Y_{obs}$, i.e., there is no missing data. Methods for handling missing data and one stage of partial synthesis simultaneously are presented by Reiter (2004b).

The agency generates the partially synthetic data in two stages. Let $Y_a^{(i)}$ be the values imputed in the first stage in nest i, where $i = 1, \ldots, m$. Let $Y_b^{(i,j)}$ be the values imputed in the second stage in data set j in nest i, where $j = 1, \ldots, r$. Let Y_{nrep} be the values of Y_{obs} that are not replaced with synthetic data and hence are released as is. Let $Z_{a,l} = 1$ if unit l, for $l = 1, \ldots, s$, is selected to have any of its first-stage data replaced, and let $Z_{a,l} = 0$ otherwise. Let $Z_{b,l}$ be defined similarly for the second-stage values. Let $Z = (Z_{a,1}, \ldots, Z_{a,s}, Z_{b,1}, \ldots, Z_{b,s})$.

To create $Y_a^{(i)}$ for those records with $Z_{a,l} = 1$, first the agency draws from $f(Y_a \mid D_{obs}, Z)$, conditioning only on values not in Y_b. Second, in each nest, the agency generates $Y_b^{(i,j)}$ for those records with $Z_{b,l} = 1$ by drawing from $f(Y_b^{(i,j)} \mid D_{obs}, Z, Y_a^{(i)})$. Each synthetic data set $D^{(i,j)} = (X, Y_a^{(i)}, Y_b^{(i,j)}, Y_{nrep}, I, Z)$. The entire collection of $M = mr$ data sets, $D_{syn} = \{D^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, r\}$, with labels indicating the nests, is released to the public.

To obtain inferences from nested partially synthetic data, we assume the analyst acts as if each $D^{(i,j)}$ is a sample according to the original design. We require the integral

$$f(Q|D_{syn}) = \int f(Q|D_{obs}, D_{syn}) f(D_{obs}|D_{syn}) dD_{obs}. \tag{3.2}$$

Unlike in fully synthetic data, there is no intermediate step of completing populations. This integral can be approximated numerically using the approach described in the Supplement. Here we present a straightforward approximation. For large m and r, we approximate (3.2) with a normal distribution with $E(Q|D_{syn}) = \bar{q}_M$ and variance $T_p = \bar{u}_M + b_M/m$. For small m and r, we can

use a t-distribution for inferences, $(\bar{q}_M - Q) \sim t_{\nu_p}(0, T_p)$. The degrees of freedom $\nu_p = (m-1)(1+m\bar{u}_M/b_M)^2$. This is derived by matching the first two moments of T_p to a chi-squared distribution with ν_p degrees of freedom, as shown in the Supplement. We note that $T_p > 0$ always, so that negative variance estimates do not arise in two-stage partial synthesis.

4. Illustrative Simulations

Given D_{syn}, analysts can use Monte Carlo methods to approximate $f(Q|D_{syn})$. However, as with other multiple imputation settings, many would prefer to use the simpler combining rules presented in Section 3. Therefore, it is important to evaluate the frequentist properties of these methods, which we do with simulation studies. We simulate from correct predictive distributions in these studies to focus on the properties of the approximations. Of course, for genuine data the larger issue is the validity of the synthesis models themselves. We discuss this further in Section 5.

We generated a population of $N = 100,000$ records comprising five variables, $Y_1, \ldots, Y_5$. The (Y_1, Y_2) were drawn from a joint t-distribution with 20 degrees of freedom and a correlation of 0.5. The (Y_3, Y_4, Y_5) were drawn from the normal distribution $N(\mu, \Sigma)$, where $\mu_1 = 1.5Y_1 + 1.5Y_2$, $\mu_1 = 2.5Y_1 + 2.5Y_2$, $\mu_3 = -3.0Y_1 - 3.0Y_2$, and Σ has variance elements equal to 30 and covariance elements equal to 15. The observed data, D_{obs}, comprised the values of $(Y_1, \ldots, Y_5)$ for a simple random sample of $s = 1,000$ records from this population. We repeated the simulation 5,000 times for both partial and full synthesis, each time drawing a new D_{obs} from the population.

We estimate five quantities: the population mean of Y_3 ($\bar{Y}_3$), the regression coefficients of Y_1 (β_1) and of Y_5 (β_5) in a regression of Y_3 on all other variables, and the regression coefficients of Y_2 (α_2) and of Y_5 (α_5) in a regression of Y_1 on all other variables. For simplicity, we do not use finite population correction factors when computing the $u^{(i,j)}$.

4.1. Results for partial synthesis

For the partial synthesis simulation, Y_1 and Y_2 were fixed at their original values. We treated $Y_a = (Y_3, Y_4)$ as the first stage variables and $Y_b = Y_5$ as the second stage variable. For each synthetic data set $D^{(i,j)}$, where $i = 1, \ldots, m$ and $j = 1, \ldots, r$, we generated $Y_a^{(i)}$ by sampling from $f(Y_3, Y_4|D_{obs})$, and we simulated $Y_b^{(i,j)}$ by sampling from $f(Y_5|D_{obs}, Y_a^{(i)})$, with noninformative prior distributions on all parameters. The released data comprised the mr copies of $(Y_a^{(i)}, Y_b^{(i,j)})$.

Table 1 summarizes the results for several combinations of (m, r). The averages of the $\bar{q}_M$ across the iterations are within simulation error of their corresponding population values; we do not report them in the table. For most

Table 1. Simulation results for two-stage partially synthetic data.

				95% CI Cov.	
(m, r)	Q	Var$(\bar{q}_M)$	Avg. T_p	Synthetic	Observed
3, 3	$\bar{Y}_3$	0.0588	0.0572	94.0	95.2
	β_1	0.0648	0.0666	95.2	95.1
	β_5	0.00115	0.00116	95.0	95.0
	α_2	0.00118	0.00109	93.9	93.6
	α_5	0.0000165	0.0000156	94.3	94.4
5, 5	$\bar{Y}_3$	0.0499	0.0494	95.1	94.9
	β_1	0.0553	0.0565	94.9	95.1
	β_5	0.00103	0.00102	94.7	94.9
	α_2	0.00108	0.00101	94.4	94.4
	α_5	0.0000151	0.0000141	94.3	94.3
5, 20	$\bar{Y}_3$	0.0471	0.0494	95.9	95.6
	β_1	0.0560	0.0554	94.6	95.0
	β_5	0.000955	0.000972	95.2	94.9
	α_2	0.00106	0.000989	93.9	94.0
	α_5	0.0000146	0.0000137	94.2	94.0
20, 5	$\bar{Y}_3$	0.0391	0.0404	95.6	95.1
	β_1	0.0474	0.0472	94.9	94.7
	β_5	0.000917	0.000921	94.7	94.8
	α_2	0.00107	0.000974	93.5	93.6
	α_5	0.0000142	0.0000132	94.4	94.7
20, 20	$\bar{Y}_3$	0.0396	0.0403	95.3	95.2
	β_1	0.0459	0.0470	95.4	95.2
	β_5	0.000879	0.000911	95.3	95.2
	α_2	0.00104	0.000968	94.4	94.0
	α_5	0.0000141	0.0000131	93.9	93.7

estimands, T_p is nearly unbiased for $Var(\bar{q}_M)$. The coverage rates for the 95% confidence intervals based on the methods in Section 3.2 are within simulation error of those based on D_{obs}. The methods have good frequentist properties in this simulation.

4.2. Results for full synthesis

For the full synthesis simulation, we assumed that (Y_1, Y_2) were known for all N records and that (Y_3, Y_4, Y_5) were known only for the s sampled records. Using an analogy with the IAB Establishment Panel synthesis, the (Y_1, Y_2) are like variables found in the German Social Security Data; the (Y_3, Y_4, Y_5) are

like variables only found in the Establishment Panel; and, concatenating all five variables for the s records is like matching the information from the GSSD for the Establishment Panel respondents. For simplicity, we did not use stratified sampling.

We treated $Y_a = (Y_1, Y_2)$ as the first stage variables and $Y_b = (Y_3, Y_4, Y_5)$ as the second stage variables. For each synthetic data set $D^{(i,j)}$, where $i = 1, \ldots, m$ and $j = 1, \ldots, r$, we generated $Y_a^{(i)}$ by taking a random sample of $n_{syn} = 1,000$ records from the population and using their values of (Y_1, Y_2). We generated $Y_b^{(i,j)}$ for these records by sampling from $f(Y_3, Y_4, Y_5 | D_{obs}, Y_a^{(i)})$, with noninformative prior distributions on all parameters. The released data comprise the mr copies of the $(Y_a^{(i)}, Y_b^{(i,j)})$. By including the imputations for the first stage variables in the released data, we deviate from the IAB Establishment Panel synthesis. However, this enables evaluations of relationships between variables imputed at different stages.

Table 2 summarizes the results for several combinations of m and r. The averages of $\bar{q}_M$ across the iterations are again within simulation error of their corresponding population values and not reported. For most estimands, T_f is nearly unbiased for $Var(\bar{q}_M)$. For $m = r = 3$, the values of T_f are frequently negative. This results from high variability in b_M and $\bar{w}_M$. Negative variance estimates become less frequent as M increases, since the variability in b_M and $\bar{w}_M$ decreases. The always positive variance estimator T_f^* is, as expected, conservative.

The column labeled "95% CI Cov*" displays the coverage rates of synthetic 95% confidence intervals based on T_f^* and on the t-distributions with ν_f^* degrees of freedom. When m or r is small, the intervals have greater than nominal coverage rates. This is primarily due to the conservatism of T_f^*. It also results from small values of ν_f^*, sometimes less than one, that arise because of inadequacies in the approximations for modest m and r. To avoid unrealistically small values, we tried a modified degrees of freedom, $\nu_f^{**} = \max\{(m-1), \nu_f^*\}$. As displayed in the column labeled "95% CI Cov**," this results in coverage rates closer to 95%. The adjustments used to obtain T_f^* and ν_f^{**} are somewhat *ad hoc*, and the properties of these simple fixes need to be studied further. We note that confidence intervals based on normal distributions for all iterations led to consistently lower than nominal coverage rates.

We also examined the variance estimator for one-stage fully synthetic data developed by Raghunathan et al. (2003). That is, we ignored the nesting. The one-stage variance estimator tends to underestimate variances. This underestimation becomes less severe as m and r increase.

Table 2. Simulation results for two-stage fully synthetic data.

(m, r)	Q	Var($\bar{q}_M$)	Avg. T_f	$\%T_f < 0$	Avg. T_f^*	95% CI Cov*	95% CI Cov**
3, 3	$\bar{Y}_3$	0.0409	0.0389	15.7	0.0448	97.6	95.2
	β_1	0.0537	0.0533	12.3	0.0587	98.0	95.9
	β_5	0.00108	0.00106	12.2	0.00117	98.0	96.2
	α_2	0.000766	0.000850	24.8	0.00109	97.6	96.3
	α_5	0.0000121	0.0000126	19.3	0.0000151	97.8	95.7
5, 5	$\bar{Y}_3$	0.0327	0.0335	3.6	0.0349	99.2	95.5
	β_1	0.0458	0.0471	1.8	0.0479	98.8	96.0
	β_5	0.000929	0.000942	1.8	0.000958	98.8	95.8
	α_2	0.000615	0.000686	12.1	0.000802	99.6	95.0
	α_5	0.00000980	0.0000109	6.0	0.0000116	99.6	95.6
5, 20	$\bar{Y}_3$	0.0319	0.0319	0.0	0.0319	95.6	95.4
	β_1	0.0448	0.0449	0.0	0.0449	95.4	95.4
	β_5	0.000878	0.000901	0.0	0.000901	95.8	95.7
	α_2	0.000581	0.000662	4.1	0.000701	99.1	95.0
	α_5	0.00000925	0.0000103	0.4	0.0000103	97.3	96.0
20, 5	$\bar{Y}_3$	0.0303	0.0308	0.0	0.0308	95.9	94.8
	β_1	0.0454	0.0450	0.0	0.0450	95.1	94.9
	β_5	0.000885	0.000890	0.0	0.000890	95.1	94.8
	α_2	0.000501	0.000576	0.7	0.000582	98.4	94.1
	α_5	0.00000870	0.0000953	0.1	0.00000955	96.9	95.0
20, 20	$\bar{Y}_3$	.0312	.0305	0.0	0.0305	94.6	94.6
	β_1	0.0426	0.0444	0.0	0.0444	95.5	95.5
	β_5	0.000850	0.000885	0.0	0.000885	95.6	95.6
	α_2	0.000492	0.000573	0.0	0.000573	96.6	95.9
	α_5	0.00000869	0.00000946	0.0	0.00000946	96.0	96.0

5. Concluding Remarks

The key to any synthetic data approach is the imputation models. For full synthesis or partial synthesis with high fractions of replacement, the validity of inferences depends critically on the validity of the models used to generate the synthetic data. When the models fail to reflect certain relationships accurately, analysts' inferences also do not reflect those relationships. Similarly, incorrect distributional assumptions built into the models are passed on to users' analyses. On the other hand, for partial synthesis that replaces only a modest fraction of values and leaves many original values on the file, inferences are less sensitive to the assumptions of the imputation models.

Agencies need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses, especially with high fractions

of synthesis. For example, agencies might include the code for synthetic data generation with public releases of data. Or, they might include generic statements that describe the imputation models, such as "Main effects for age, sex, and race are included in the imputation models for education." Analysts who desire finer detail than afforded by the imputations may have to apply for special access to the observed data.

Many analysts of public use data files estimate domain means and basic regressions, whereas agencies generate imputations from more complicated models. Such mismatches have been termed uncongeniality in the literature on multiple imputation for missing data (Meng (1994, 2002)). There has been little theoretical work on the consequences of uncongeneality for synthetic data. However, frequentist evaluations based on simulated data (Reiter (2002, 2003)) suggest that one-stage synthetic data inferences have good properties–in the sense that coverage rates of confidence intervals are near or exceed nominal rates–when the analysts' inferences can be embedded in the imputation models. The simulation results in Section 4 are in accord with these findings. Empirical investigations of frequentist properties based on genuine data tell basically the same story (Reiter (2005b,d)).

For any particular setting, the agency can evaluate the synthetic data models by comparing inferences made with synthetic data to those made with observed data. Thus, evaluating imputation models for disclosure limitation is conceptually more straightforward than evaluating imputation models for missing data (Reiter (2004b)).

The most extensive testing of the analytical validity of synthetic data has been done for the Survey of Income and Program Participation (SIPP). In 2001, the Census Bureau, the Internal Revenue Service, and the Social Security Administration decided to supplement the information on SIPP panels from 1990 - 1996 with detailed earnings and Social Security benefits histories. The three agencies agreed to release a version of the linked data only if all but four out of over 600 variables were synthesized. To evaluate the synthesis, Abowd, Stinson and Benedetto (2006) compared inferences from the observed and synthetic data for a large number of estimands. The synthesis models reproduced the observed data univariate distributions for all variables but the highly skewed wealth-related variables, and resulted in synthetic data confidence intervals very similar to the corresponding observed data intervals for summary statistics for important earnings and benefit measures (e.g., work histories, annual earnings) for all major demographic subgroups. The models had mixed success with linear and logistic regression coefficients: synthetic and observed data confidence intervals were similar for some coefficients but not for others.

Such empirical evidence aside, some inferences will deteriorate significantly because of imperfect imputation models. When simulating high fractions of data,

even small biases can cause substantial reductions in frequentist validity. These biases may be hard to detect from any meta-data released by the agency describing the synthesis process. For this reason, it is arguably essential that agencies develop ways to provide feedback to users about the quality of the synthetic data inferences for specific estimands. One possibility is to build a verification server, as suggested by Reiter, Oganian and Karr (2009). The basic idea is as follows. The data user performs an analysis of the synthetic data, using whatever software she wishes. She then submits a description of the analysis to the verification server; for example, regress attribute 5 on attributes 1, 2, 4 and the logarithm of attribute 6. The verification server performs the analysis on both the confidential and synthetic data, and from the results calculates analysis-specific measures of the fidelity of the one to the other. For example, for any regression coefficient, measure the overlap in its confidence intervals (Karr et al. (2006)) computed from the confidential and synthetic data. The verification server returns the value of the fidelity measure to the user. If the user feels that the intervals overlap adequately, the synthetic data have high utility for their analysis. With such feedback, analysts can avoid publishing–in the broad sense–results with poor quality, and be confident about results with good quality.

Verification servers are not a panacea. As illustrated by Reiter et al. (2009), fidelity measures provide intruders with information about the real data, albeit in a convoluted form, that could be used for disclosure attacks. It may be possible to blunt these attacks by providing coarse fidelity measures or by limiting the types of queries that the server answers. Assessing and reducing the risks of providing fidelity measures are topics of ongoing research.

Additional topics for future research specific to two-stage synthesis include methods for selecting m and r based on risk-utility evaluations, for using the M data sets to do significance tests of multi-component hypotheses and other multivariate inference, and for handling missing data and confidentiality simultaneously, perhaps in a three-stage imputation procedure.

For many data sets, concerns over confidentiality make it nearly impossible to release public use data as is. As resources available to malicious data users attempting re-identifications continue to expand, the alterations needed to protect data with traditional disclosure limitation techniques–such as swapping, adding noise, or microaggregation–may become so extreme that, for many analyses, the released data are no longer useful. Synthetic data, on the other hand, have the potential to enable data dissemination while preserving data utility. By synthesizing in two stages, data producers can improve the risk-utility profile, or reduce the labor costs, of their data releases.

Acknowledgement

This research was supported by the National Science Foundation grant, ITR-0427889.

References

Abowd, J., Stinson, M. and Benedetto, G. (2006). Final report to the social security administration on the SIPP/SSA/IRS public use file project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.bls.census.gov/sipp/synth_data.html`.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (Edited by P. Doyle, J. Lane, L. Zayatz and J. Theeuwes), 215-277. North-Holland, Amsterdam.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases* (Edited by J. Domingo-Ferrer and V. Torra), 290-297. Springer-Verlag, New York.

Drechsler, J., Dundler, A., Bender, S., Rässler, S. and Zwick, T. (2008). A new approach for disclosure control in the IAB establishment panel-Multiple imputation for a better data access. *Adv. Stat. Anal.* **92**, 439-458.

Duncan, G. T., Keller-McNulty, S. A. and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Tech. rep., U.S. National Institute of Statistical Sciences.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University.

Fienberg, S. E., Makov, U. E. and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *J. Off. Statist.* **14**, 485-502.

Gomatam, S., Karr, A. F., Reiter, J. P. and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statist. Sci.* **20**, 163-177.

Harel, O. and Schafer, J. (2003). Multiple imputation in two stages. In *Proc. Fed. Com. on Statist. Methodol.* 2003.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *Amer. Statist.* **60**, 224-232.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 survey of consumer finances. In *Record Linkage Techniques*, 1997 (Edited by W. Alvey and B. Jamerson), 248-267. National Academy Press, Washington, D.C..

Little, R. J. A. (1993). Statistical analysis of masked data. *J. Off. Statist.* **9**, 407-426.

Little, R. J. A., Liu, F. and Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Persepectives* (Edited bu A. Gelman and X. L. Meng), 141-152. John Wiley, New York.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statist. Sci.* **9**, 538-558.

Meng, X. L. (2002). A congenial overview and investigation of multiple imputation inferences under uncongeneality. In *Survey Nonresponse* (Edited by R. M. Groves, D. A. Dillman, J. L. Eltinge and R. Little), 343-356. John Wiley, New York.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In *Privacy in Statistical Databases* (Edited by J. Domingo-Ferrer and L. Franconi), 177-188. Springer-Verlag, New York.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Surv. Methodol.* **27**, 85-96.

Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Statist.* **19**, 1-16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *J. Off. Statist* **18**, 531-544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Surv. Methodol.* **29**, 181-189.

Reiter, J. P. (2004a). New approaches to data dissemintation: A glimpse into the future (?). *Chance* **17**, 12-16.

Reiter, J. P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235-242.

Reiter, J. P. (2005a). Estimating identification risks in microdata. *J. Amer. Statist. Assoc.* **100**, 1103-1113.

Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *J. Roy. Statist. Soc. Ser. A* **168**, 185-205.

Reiter, J. P. (2005c). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *J. Statist. Plann. Inference* **131**, 365-377.

Reiter, J. P. (2005d). Using CART to generate partially synthetic, public use microdata. *J. Off. Statist.* **21**, 441-462.

Reiter, J. P., Oganian, A. and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Comput. Statist. Data Anal.* **53**, 1475-1482.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *J. Off. Statist.* **9**, 462-468.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statist. Neerlandica* **57**, 3-18.

Shen, Z. (2000). Nested multiple imputation. Ph.D. thesis, Harvard University, Dept. of Statistics.

Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251, U.S.A.

E-mail: jerry@stat.duke.edu

Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany.

E-mail: joerg.drechsler@iab.de.

(Received August 2007; accepted December 2008)